International Journal of Geo-Information

MDPI

*Article*

# RetiGlue-SLAM: A Robust Visual SLAM System Designed for Low-Light and Complex Environments

**Yiming Li [1,2], Xianpu Liang [1,2], Guangming Guo [3], Peng Su [1,2,*] and Qi An [4]**

[1] Beijing Key Laboratory of Intelligent Operation and Maintenance and Control of High-End Equipment, Beijing Information Science and Technology University, Beijing 100192, China; liyimingxf@bistu.edu.cn (Y.L.); 2024020127@bistu.edu.cn (X.L.)
[2] Mechanical Electrical Engineering School, Beijing Information Science & Technology University, Beijing 100192, China
[3] Intelligent Equipment Research Institute, Beijing Academy of Science and Technology, Beijing 100061, China
[4] State Key Laboratory of Tribology, Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China
* Correspondence: supeng@bistu.edu.cn

## Highlights

- RetiGlue-SLAM couples Retinexformer, SuperPoint, and LightGlue with the ORB-SLAM3 geometric back end for low-light visual-inertial SLAM.
- An ablation study on EuRoC isolates the learned front end from the full enhancement-and-matching pipeline.
- The prototype front end requires 41 ms per stereo frame (24.39 FPS); end-to-end SLAM latency is reported as a limitation.
- The method improves tracking completion on the night-time 4Seasons sequence while retaining competitive accuracy on normal-light EuRoC sequences.

## Abstract

Visual simultaneous localization and mapping (SLAM) often degrades in low-light environments because of insufficient brightness, reduced contrast, and loss of texture, which cause sparse features, accumulated mismatches, and unstable trajectories. This paper presents RetiGlue-SLAM, a visual-inertial SLAM framework that retains the third version of ORB-SLAM (ORB-SLAM3) as the system backbone and improves the front end through task-driven coupling of three components: Retinexformer for low-light image enhancement, SuperPoint for learned local feature extraction, and LightGlue for robust cross-frame matching. The enhanced images and learned correspondences are integrated into the tracking thread, whereas the geometric back-end and inertial optimization of ORB-SLAM3 remain unchanged. Experiments on low-light sequences from EuRoC and the night-time Old Town sequence of 4Seasons show improved feature matching, tracking completion, and localization accuracy relative to the baseline. A two-stage ablation separates the learned front end from the full RetiGlue-SLAM pipeline, and a module-level runtime analysis reports a front-end latency of 41 ms per stereo frame (24.39 FPS). The results support RetiGlue-SLAM as a benchmarked low-light visual-inertial localization framework; field deployment and end-to-end real-time operation remain outside the scope of this study.

**Keywords:** low-light environment; visual SLAM; image enhancement; deep learning; feature extraction and matching; visual-inertial odometry

Check for updates

Academic Editors: Wolfgang Kainz, Eliseo Clementini and Zhiyong Zhou

Received: 27 August 2026
Revised: 4 October 2026
Accepted: 10 October 2026
Published: 14 October 2026

**Copyright:** © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

In recent years, simultaneous localization and mapping (SLAM) has become a key technology for autonomous navigation of mobile robots and environmental perception in complex scenes. Among various approaches, visual SLAM has emerged as an important research direction because of its low cost, high information density, lightweight system design, and flexible deployment [1,2]. Although multisensor perception systems demonstrate strong capabilities in complex scenes [3], they often suffer from high system integration complexity, large hardware overhead, and high deployment costs, which limit their application on lightweight platforms.

In typical scenarios such as night-time reconnaissance, underground search, urban streets, and poorly lit roads, the environment often presents challenges, including GNSS denial, insufficient illumination, texture degradation, and increased dynamic interference. These factors impose significant limitations on conventional visual SLAM in terms of feature extraction, data association, and pose estimation. To address these issues, existing research has mainly followed three technical approaches: traditional methods based on handcrafted features and geometric optimization; learning-based methods that directly estimate poses from input data; and hybrid approaches that retain the classical geometric framework while incorporating learning-based modules to enhance front-end perception and matching capabilities. Overall, the introduction of learning-based techniques has significantly expanded the adaptability of visual SLAM to complex illumination conditions, although differences remain among these methods in terms of robustness, real-time performance, and generalization capability.

Currently, among traditional visual SLAM methods, those based on manually designed features still dominate the field [4,5]. ORB-SLAM [6], proposed by Mur-Artal et al., introduced a loop-closure module on the basis of PTAM and consistently used ORB features across its three parallel threads—tracking, mapping, and loop closing—thereby establishing a relatively complete feature point-based visual SLAM framework. Subsequently, the same research team developed ORB-SLAM2 [7] and ORB-SLAM3 [8], extending the applicability of feature-based visual SLAM to a wide range of configurations, including monocular, stereo, RGB D, and visual-inertial setups. Owing to its ability to run in real time on CPUs and its strong robustness, the ORB SLAM series has become a classic approach in the field of visual SLAM. However, in scenes with weak texture or significant illumination changes, the number and quality of extracted feature points drop markedly, which impairs the system's matching and tracking capabilities and adversely affects its overall accuracy and stability.

With the continued advancement of deep learning in image recognition and feature representation [9], researchers have begun to introduce it into visual localization and SLAM tasks. Existing studies can be roughly divided into two main approaches. The first is the learning-based end-to-end approach, which directly uses deep neural networks to estimate camera motion and scene state from images or multimodal inputs [10,11]. Representative works include UnDeepVO [12], DeepVO [13], and DROID-SLAM [14]. Such methods can overcome some limitations of manually designed features and directly learn nonlinear mappings between complex appearance changes and motion states. However, their performance depends on training-data coverage, and they still face challenges in cross-scene generalization, complex illumination, interpretability, and real-time deployment. The second approach is the hybrid approach, which retains the main framework of geometric SLAM while using learning-based techniques to enhance key modules such as image enhancement, feature extraction, feature matching, or scene understanding [15–17]. Among these, the SuperPoint-based method for learning local features [15] has been shown to

provide more stable keypoint and descriptor representations than traditional manually designed features under complex illumination conditions.

Although learning-enhanced front ends offer new possibilities for improving SLAM robustness in complex illumination environments, they still exhibit notable shortcomings in low-light scenarios. On one hand, simply replacing traditional front ends with learning-based features does not fundamentally address the lack of observable information in low-light images themselves [15,18,19]; if the input images suffer from severe underexposure, strong noise, or local exposure imbalances, the feature extraction network remains significantly affected. On the other hand, applying conventional enhancement methods alone as a preprocessing step for low-light images can improve brightness to some extent, but often introduces side effects such as noise amplification, color distortion, or artificial textures [17,19], which in turn interfere with subsequent geometric estimation. Furthermore, over-reliance on end-to-end learning may weaken the interpretability and stability provided by classical geometric constraints. Therefore, how to achieve cooperative optimization among image enhancement, learning-based feature extraction, and robust matching while preserving the advantages of the geometric back end remains a key challenge in current visual SLAM research.

To address the above issues, this paper proposes an improved visual-inertial SLAM framework tailored for low-light and complex environments. The framework adopts ORB-SLAM3 as its overall architecture. At the image input stage, Retinexformer is introduced to perform brightness recovery, noise suppression, and local illumination correction on low-light images, thereby enhancing visible details in dark regions and improving front-end perception quality. At the feature level, SuperPoint is employed for keypoint detection and descriptor generation, combined with LightGlue for efficient cross-frame matching, which strengthens feature association stability under conditions of weak texture, low contrast, and sudden illumination changes. At the system level, the enhanced images, learned features, and the classical SLAM back-end optimization process are uniformly integrated to improve continuous tracking capability, localization accuracy, and mapping stability in indoor and outdoor night-time environments. Experimental results demonstrate that the proposed method effectively alleviates the localization instability and mapping failures commonly caused by illumination degradation in traditional visual SLAM under low-light conditions, exhibiting superior robustness and adaptability.

The main contributions of this paper are as follows:

1. A task-driven low-light front end. RetiGlue-SLAM retains ORB-SLAM3 as the complete visual-inertial backbone and modifies only the observation pipeline through Retinexformer-based enhancement, SuperPoint-based feature extraction, and LightGlue-based matching. The contribution is the system-level coupling of these modules under geometric constraints, not the invention of the individual networks.
2. Module coupling and implementation transparency. We specify how enhanced images, learned keypoints and descriptors, consecutive-frame matches, and the unchanged geometric back end are connected. We also report the processing order and module-level front-end latency.
3. Low-light benchmark evaluation with controlled ablation. We evaluate the method on low-light EuRoC sequences and the night-time Old Town sequence of 4Seasons, retain normal-light EuRoC sequences as controls, and add a two-stage ablation that separates the learned front end from the full RetiGlue-SLAM pipeline.
4. Scope-aware analysis. We discuss where the method is supported by the reported data and where it is not, including the absence of a field-deployment dataset, the lack of a full factorial component study, and the difference between front-end throughput

and end-to-end real-time operation. We also compare the design with recent low-light-oriented SLAM systems at the methodological level.

## 2. Related Work

### 2.1. Traditional Feature-Based Visual SLAM Methods

Feature-based methods are the most mature and widely used class of techniques in visual SLAM. Such methods typically extract salient local features from images and use feature matching to estimate camera motion and update the map. Commonly used manually designed features include Harris corners [20], SIFT [21], SURF [22], FAST [23], and ORB [24]. These methods achieve good efficiency and geometric interpretability in scenes with rich texture, stable imaging conditions, and relatively consistent illumination, and thus have long occupied a dominant position in visual SLAM research. MonoSLAM, proposed by Davison et al., was a pioneering work that achieved real-time monocular localization and mapping. However, it lacked global optimization and loop-closure capabilities, and its computational complexity increased significantly with the scale of the scene. Subsequently, PTAM, introduced by Klein and Murray, separated tracking and mapping into parallel threads, greatly improving system efficiency and laying the foundation for the development of subsequent feature-based SLAM systems. RTAB MAP [25], proposed by Labbé and Michaud, achieved strong performance in loop closure detection and long-term mapping, making it particularly suitable for large-scale online scenarios. ORB SLAM and ORB SLAM2, developed by Mur Artal and Tardós, further advanced the systematic development of feature point-based methods. By leveraging ORB features, a relocalization mechanism, and loop-closure optimization, they constructed a relatively complete real-time visual SLAM system that supports multiple configurations, including monocular, stereo, and RGB D, and has been widely applied in both indoor and outdoor environments.

However, the performance of such methods heavily depends on the quality of the input images, and low-light environments fundamentally degrade their front-end perception conditions. Unlike general complex scenes, low-light conditions not only cause overall image dimming but also introduce issues such as reduced contrast, increased noise, uneven local exposure, and loss of structural details in dark areas. These degradations directly affect the response stability of manually designed features, leading to a decrease in the number of detectable keypoints, deterioration of their spatial distribution, and reduced descriptor discriminability. Ultimately, this results in insufficient constraints for cross-frame matching and, in severe cases, can cause tracking failures or even mapping breakdown. Therefore, the stable performance of traditional feature-based methods in standard environments does not imply that they can naturally adapt to night-time, poorly lit, or dramatically changing illumination conditions.

To address this issue, recent studies have begun incorporating image enhancement strategies into the front end of traditional feature-based SLAM to improve feature extractability in low-light images. For instance, Fang et al. proposed HE SLAM [26], which attempts to extract more effective ORB features from low-contrast images using histogram equalization. However, under significant illumination changes, manually designed features still struggle to maintain sufficient robustness. Thus, while traditional feature-based methods offer advantages such as clear geometric constraints and well-established implementation, their adaptability in complex low-light scenarios remains limited.

### 2.2. End-to-End Learning-Based Methods

With the successful application of deep learning in computer vision tasks, researchers have begun exploring approaches that directly learn the mapping between camera motion and scene structure from image sequences. DeepVO, which combines convolutional net-

works with recurrent networks to directly estimate camera pose from consecutive RGB images, is one of the early learning-based end-to-end visual odometry methods. UnDeepVO further proposed an unsupervised monocular visual odometry framework that jointly learns depth and pose without requiring ground truth pose annotations. Subsequently, methods such as DROID SLAM have integrated learning-based networks more closely with optimization processes, further improving the performance of learning-based visual localization.

For complex illumination environments, some studies have attempted to combine low-light enhancement and learning-based pose estimation into a unified framework. VIOLearner [27] integrates visual and inertial information and improves motion estimation performance through multi-scale error correction. PoseConvGRU [28] employs a recurrent convolutional structure to learn geometric relationships between frames. PU-PoseNet [29] introduces a prediction-update mechanism to enhance pose estimation stability. More directly targeting low-light environments, ULL-SLAM [30] incorporates a low-light enhancement branch and a self-supervised feature extraction branch in the front end to improve matching quality under weak lighting conditions. In addition, Parisetto et al. [31] proposed an end-to-end system that combines attention mechanisms and neural graph optimization to replace traditional LSTM layers for drift correction, achieving better performance than DeepVO in simulated environments.

Overall, end-to-end methods can learn nonlinear appearance-to-motion mappings, but they require large training sets, are costly to train and deploy, and remain difficult to interpret or generalize across sites. They therefore complement, rather than replace, geometry-based systems.

### *2.3. Learning-Based Enhanced SLAM Methods*

Compared with fully end-to-end visual localization methods, approaches that retain the traditional geometric optimization framework while using learning-based techniques to enhance front-end perception and data association are better aligned with the practical demands of visual SLAM in low-light environments. The fundamental reason is that the main difficulties in low-light scenes originate from degraded image quality—such as insufficient brightness, uneven local exposure, noise amplification, reduced contrast, and loss of structural information in dark regions. These issues directly impair keypoint detection, feature description, and matching quality across frames, thereby affecting the stability of subsequent pose estimation. Therefore, existing research tends to introduce learning-based modules to improve front-end perception under low-light conditions while preserving the stability and interpretability of the geometric back end, rather than relying entirely on end-to-end regression to replace the entire SLAM pipeline.

In the design of learning-based front ends, existing work has primarily focused on improving traditional SLAM systems through learning-based feature extraction, depth estimation, and self-supervised matching. LIFT-SLAM [32] enhances matching quality in monocular SLAM by incorporating learned local features. CNN-SLAM [33] uses depth prediction to mitigate geometric degradation in low-light and low-texture regions. The graph convolutional network (GCN) proposed by Tang et al. and its subsequent extensions [34,35] improve front-end representation capabilities by jointly learning keypoints and descriptors. SP-Flow [36] attempts to replace the traditional ORB front end with self-supervised optical flow correspondences. These studies demonstrate that, compared with manually designed features, learned features generally exhibit higher repeatability and discriminability in scenarios with weak texture, viewpoint changes, and low-light conditions, thus offering greater potential to improve tracking and matching performance under challenging environments.

Replacing handcrafted features with learned features alone does not fully solve low-light localization, because severe underexposure, noise, and uneven illumination degrade the input itself. A learned detector can still fail when the image contains too little stable structure.

Therefore, in recent years, a more targeted research trend has emerged that combines low-light image enhancement with learning-based front ends. By first improving the observability of low-light images before performing keypoint extraction and cross-frame matching, this approach enhances the ability of visual localization and SLAM systems to operate continuously in night-time and poorly lit scenes. Relevant studies indicate that the difficulties in low-light environments are not limited to reduced brightness but also include increased noise, uneven local illumination, and loss of structural details. Therefore, the enhancement process needs to balance brightness recovery, suppression of degradation, and preservation of structural information.

Retinexformer [37] provides a more effective implementation for low-light image enhancement by unifying illumination estimation and degradation restoration into a single-stage framework. Matching in the Dark [38] further highlights that low-light image matching is a critical bottleneck affecting real-time recovery of image structure and visual SLAM performance. Systematic experiments in Twilight SLAM [39] demonstrate that task-driven collaborative design between low-light enhancers and the SLAM front end is essential. In addition, research on Retinex enhancement and learned feature optimization for SLAM in low-light environments [40] has validated the effectiveness of combining image enhancement, feature extraction, and robust matching in improving localization accuracy and stability in low-light scenes. It is thus evident that the synergistic integration of low-light enhancement, learned local features, and geometric optimization frameworks has become an important research direction for enhancing visual SLAM performance in low-light conditions.

Hybrid SLAM systems generally improve robustness under weak texture and changing illumination, but their performance depends on the training distribution and on the balance between learned components and geometric constraints. This balance is the central design issue addressed in this work.

In summary, the most credible direction for low-light SLAM is not to replace the entire geometric pipeline, but to improve the image observations and feature correspondences supplied to it. RetiGlue-SLAM follows this principle and retains the ORB-SLAM3 back end.

## 3. Materials and Methods

### *3.1. System Framework*

The overall framework is shown in Figure 1. RetiGlue-SLAM uses ORB-SLAM3 as the complete system backbone, including its tracking, local mapping, loop-closure, and nonlinear optimization modules. The proposed changes are confined to the front-end observation pipeline. Low-light input images are first processed by Retinexformer [37] to improve brightness and suppress degradation while preserving structural information. The enhanced images are then passed to SuperPoint [15], which extracts learned keypoints and descriptors, and LightGlue [41], which establishes cross-frame correspondences. The resulting matches are supplied to the original ORB-SLAM3 tracking thread together with camera and IMU data. Thus, ORB-SLAM3 provides the system architecture and geometric back end, while Retinexformer, SuperPoint, and LightGlue form the modified perception front end.

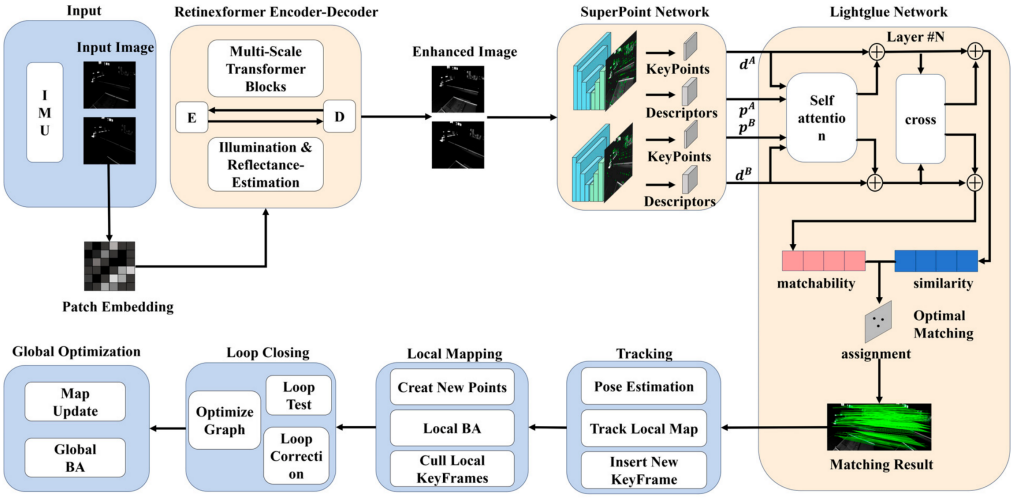



**Figure 1.** The RetiGlue-SLAM system framework.

In the local mapping stage, the system receives keyframes from the tracking thread, performs keyframe insertion and new map point generation, and uses local bundle adjustment to jointly optimize camera poses and map structure, ensuring timely map updates and local consistency. After initialization and successful keyframe tracking, the system further optimizes keyframe states by incorporating pre-integrated errors, and improves pose estimation accuracy by minimizing reprojection errors. Once optimization is completed, the system removes redundant keyframes to control map size and improve overall operational efficiency.

In the loop closure detection thread, the system first uses a bag-of-words model to identify potential loop closure keyframes. Once a valid loop closure is detected, it estimates the similarity transformation between the current keyframe and the loop keyframe, and then performs fusion of the corresponding maps and loop closure correction. Subsequently, global graph optimization is applied to further eliminate accumulated drift, and finally global bundle adjustment is performed to jointly optimize the map and trajectory, thereby ensuring global consistency of the system in large-scale scenes.

### *3.2. Low-Light Image Processing*

To improve the quality of visual observations in low-illumination scenes, existing research often introduces classic Retinex or convolutional neural network (CNN)-based Retinex enhancement methods into the SLAM front end to preprocess input images. However, such approaches often struggle to simultaneously achieve brightness recovery, noise suppression, and color preservation, and are prone to issues such as uneven illumination, noise amplification, and color distortion, which in turn affect the accuracy of front-end feature extraction and matching. To address this, this paper adopts Retinexformer as the low-light image enhancement module, which mainly consists of an illumination estimator and a degradation restorer. By performing brightness correction and degradation restoration on the input low-light images, Retinexformer produces enhanced images with clearer details, lower noise, and more uniform illumination. Based on this, this module is integrated into the front-end processing pipeline of ORB-SLAM3 to enhance the system's feature perception and tracking capability under low-light conditions.

Traditional Retinex theory holds that a low-light image can be decomposed into the element-wise product of a reflectance component and an illumination component, meaning

that the image content is primarily determined by the scene reflectance and the external illumination, which can be expressed as:

$$I = R \odot L \tag{1}$$

where $I \in \mathbb{R}^{H \times W \times 3}$ denotes the input low-light image, $R \in \mathbb{R}^{H \times W \times 3}$ is the reflectance component, $L \in \mathbb{R}^{H \times W}$ is the illumination component, and $\odot$ denotes element-wise multiplication. Retinexformer points out that the traditional model assumes the input image is free of degradation, which does not hold in real low-light scenes; therefore, it is necessary to explicitly model disturbances such as noise, artifacts, and color distortion. To describe the actual degradation in the imaging process, perturbation terms $\widehat{R}$ for reflectance and $\widehat{L}$ for illumination are introduced into (1), yielding:

$$I = \left(R + \widehat{R}\right) \odot \left(L + \widehat{L}\right) \tag{2}$$

After expansion, we obtain:

$$I = R \odot L + R \odot \widehat{L} + \widehat{R} \odot \left(L + \widehat{L}\right) \tag{3}$$

where $R \odot \widehat{L}$ reflects the exposure error and color deviation introduced during the enhancement process, and $\widehat{R} \odot \left(L + \widehat{L}\right)$ corresponds to the noise and artifacts introduced during low-light image capture and subsequently amplified during enhancement. This modeling approach is precisely what distinguishes Retinexformer from traditional Retinex decomposition methods.

To avoid numerical instability issues caused by directly estimating the illumination map $L$, Retinexformer does not solve for $L$ directly. Instead, it learns a brightening map $\overline{L}$ that satisfies

$$L \odot \overline{L} = 1 \tag{4}$$

Therefore, the brightened image can be written as

$$I_{lu} = I \odot \overline{L} = R + C \tag{5}$$

where $I_{\mathrm{lu}}$ is the brightened image, and $C$ represents the loss terms from the environment and the processing steps. Thus, the entire low-light enhancement process can be expressed as

$$(I_{lu}, F_{lu}) = \mathcal{E}(I, L_p) \tag{6}$$

$$I_{en} = \mathcal{R}(I_{lu}, F_{lu}) \tag{7}$$

where $\mathcal{E}(\cdot)$ is the illumination estimator, $\mathcal{R}(\cdot)$ is the degradation restorer, $L_p$ is the illumination prior map obtained from the channel mean of the input image, and $F_{lu}$ is the brightened feature. This structure consists of two components: illumination estimation and degradation restoration. It not only improves brightness in dark regions but also suppresses noise amplification and color shift introduced during enhancement. The two outputs, $I_{lu}$ and $F_{lu}$ are then input to obtain the final restored image $I_{en} \in \mathbb{R}^{H \times W \times 3}$.

For notation consistency, I denotes the input low-light image, R the reflectance component, L the illumination component, C the exposure and color error, N the noise and artifact term, and the symbol between R and L denotes element-wise multiplication. The illumination estimator E(.) predicts the illumination prior and brightened feature, and the degradation restorer R(.) produces the enhanced image. The same notation is used throughout Section 3.2.

Considering that convolutional neural networks have certain limitations in capturing long-range pixel dependencies, this paper adopts an Illumination-Guided Transformer (IGT) as the degradation restoration module to enhance global context modeling capability in low-light images. The overall structure of IGT is shown in Figure 2, consisting of an encoder and a decoder, where the encoder is responsible for downsampling and high-level feature extraction, and the decoder is responsible for upsampling and detail recovery. Specifically, the brightened image $I_{lu}$ output by the illumination estimator is first passed through a $3 \times 3$ convolution with a stride of 1 to match the channel dimension of the illumination feature map $F_{lu}$. Subsequently, the encoding branch progressively compresses the feature representation layer by layer. In the decoding stage, transposed convolutions with a kernel size of $2 \times 2$ and a stride of 2 are used to restore spatial resolution, thereby reducing information loss caused by downsampling. Finally, the decoding branch outputs a residual image $I_{re} \in \mathbb{R}^{H \times W \times 3}$, which is fused with $I_{lu}$ to obtain the final enhanced image $I_{\text{en}}$.

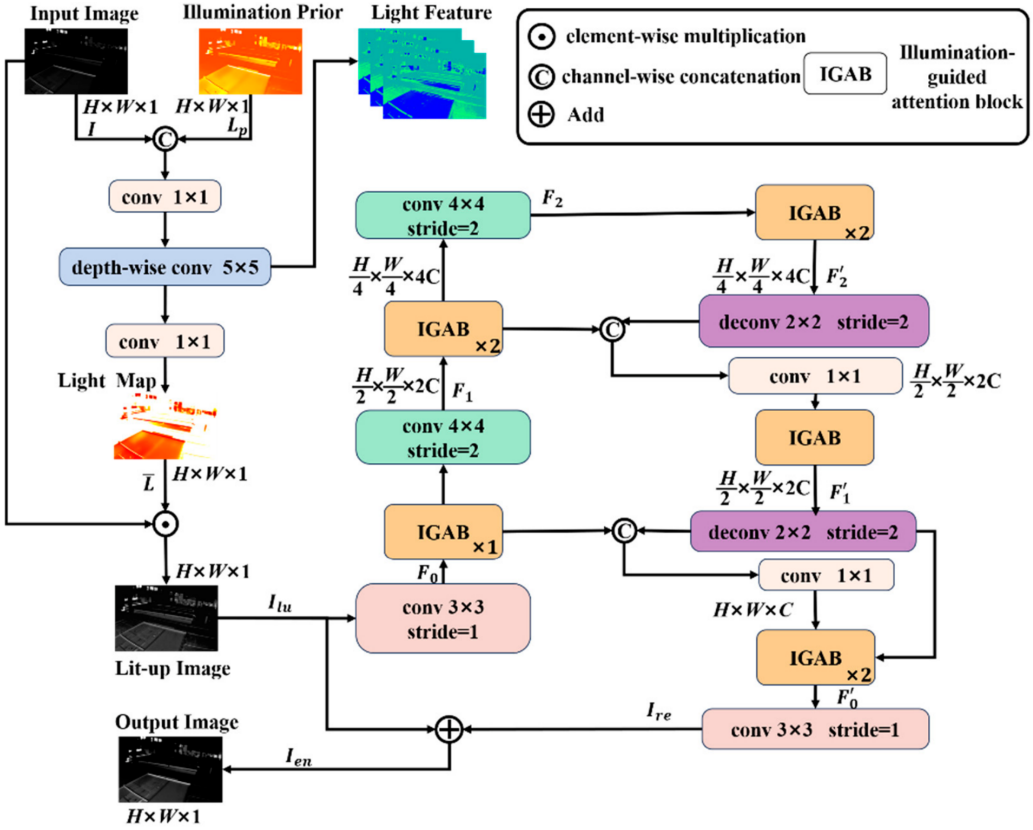



**Figure 2.** The overall structure of IGT.

In IGT, feature recovery does not rely solely on conventional convolutional operations but instead enhances information interaction across regions with different exposures by introducing an illumination-guided attention mechanism. To this end, this paper adopts the Illumination-Guided Attention Block (IGAB) as the basic building unit within the restoration module, as shown in Figure 3. Based on the network architecture design, IGAB consists of two layer-normalization (LN) layers, an illumination-guided multi-head self-attention module (IG-MSA), and a feed-forward network (FFN), with residual connections preserved outside each sub-module to enhance the stability of feature propagation. This structure improves contextual modeling capability in low-light scenes while maintaining stable network training. IGAB does not simply perform local enhancement on input features;

rather, it explicitly incorporates illumination information to achieve global dependency modeling and feature reconstruction.

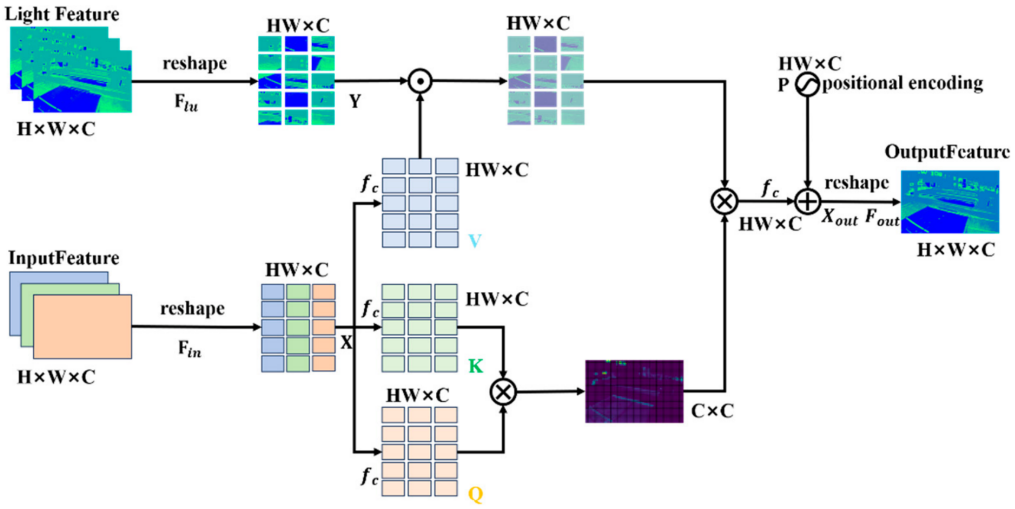



**Figure 3.** Illumination-guided attention block.

IG-MSA is the core component of Retinexformer, as shown in Figure 4. Unlike standard multi-head self-attention, which computes attention using only the input features themselves, IG-MSA introduces the illumination features Flu output by the illumination estimator as a guiding signal into the attention computation. This enhances the semantic interaction between bright regions and dark regions. Because dark regions typically suffer from more severe degradation while bright regions often retain more reliable contextual information, the illumination-guided attention can more effectively assist in the recovery of dark regions.

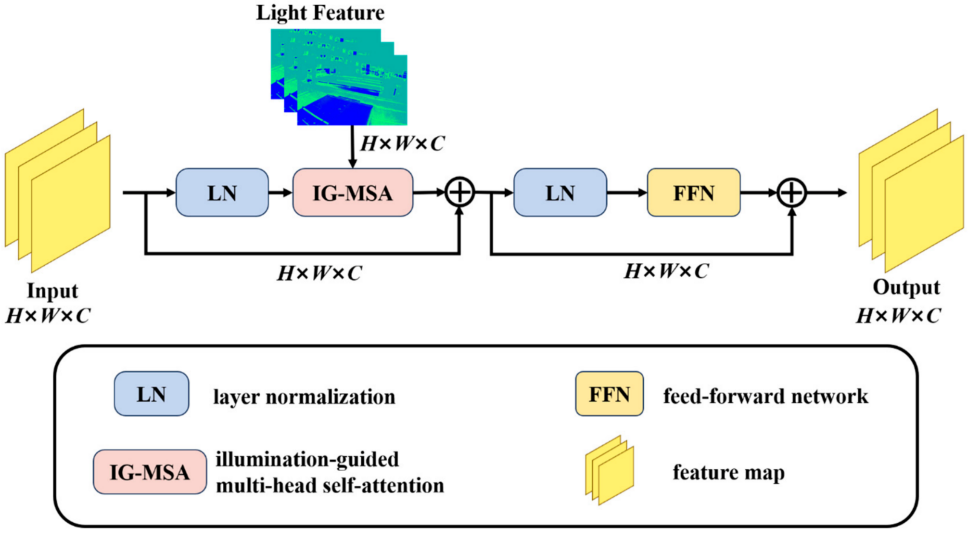



**Figure 4.** IG-MSA.

Let the input feature be $F_{in} \in \mathbb{R}^{H\times W\times C}$, which is reshaped into a token sequence $X \in \mathbb{R}^{HW\times C}$, and then divided into k heads:

$$X = [X_1, X_2, \cdots, X_k] \tag{8}$$

$$X_i \in \mathbb{R}^{HW\times d_k}, d_k = \frac{C}{k} \tag{9}$$

For the $i$-th head, its query, key, and value are, respectively:

$$Q_i = X_i W_{Q_i}^T, K_i = X_i W_{K_i}^T, V_i = X_i W_{V_i}^T \tag{10}$$

where $W_{Q_i}, W_{K_i}, W_{V_i} \in \mathbb{R}^{d_k \times d_k}$ are learnable parameters. Similarly, the illumination guidance feature $F_{lu}$ is reshaped into $Y \in \mathbb{R}^{HW \times C}$ and divided into:

$$Y = [Y_1, Y_2, \cdots, Y_k] \tag{11}$$

$$Y_i \in \mathbb{R}^{HW \times d_k} \tag{12}$$

Thus, the illumination-guided attention for the $i$-th head can be expressed as:

$$\text{Attention}(Q_i, K_i, V_i, Y_i) = (Y_i \odot V_i)\,\text{softmax}\left(\frac{K_i^T Q_i}{\alpha_i}\right) \tag{13}$$

where $\alpha_i$ is a learnable scaling parameter. Compared with conventional attention, this formulation also introduces a modulation effect of $Y_i$ on the value branch, thereby adapting the attention response to local illumination conditions. Finally, the outputs from all heads are concatenated (passed through a linear projection and positional encoding) to produce the output feature $F_{out}$.

From a computational cost perspective, the matrix multiplication operations across $k$ attention heads are a key factor affecting the efficiency of the attention module. The main computational cost of IG-MSA comes from two matrix multiplications, with its complexity given in (14), while the complexity of standard global multi-head self-attention is given in (15).

For clarity, k is the number of attention heads, C is the channel dimension, $d_k$ = C/k, and HW is the spatial resolution. These definitions are used consistently in Equations (14) and (15).

$$\mathcal{O}(\text{IG} - \text{MSA}) = k \cdot [d_k(d_k HW) + HW(d_k d_k)] = \frac{2HWC^2}{k} \tag{14}$$

$$\mathcal{O}(\text{G} - \text{MSA}) = 2(HW)^2 C \tag{15}$$

A comparison reveals that global multi-head self-attention grows quadratically with the spatial size HW, resulting in high computational cost, whereas IG-MSA grows approximately linearly with spatial size. Therefore, the IG-MSA designed in this paper reduces the computational complexity to a near-linear relationship with spatial size, as shown in (14). This allows it to be flexibly embedded across different levels of IGT, balancing modeling capability with computational efficiency, and making it suitable for repeated use in multi-scale low-light enhancement networks.

### *3.3. Learning-Based Local Feature Extraction*

Although low-light image enhancement improves the localization performance of the system to some extent, relying solely on image preprocessing remains insufficient to fully meet the requirements for robust localization in complex scenes. To further enhance the feature perception capability of the front end, this paper introduces the SuperPoint network at the feature extraction stage, as shown in Figure 5. This method adopts a fully convolutional network architecture, requiring only a single forward pass through the entire image to simultaneously perform keypoint detection and generate fixed-length descriptors.

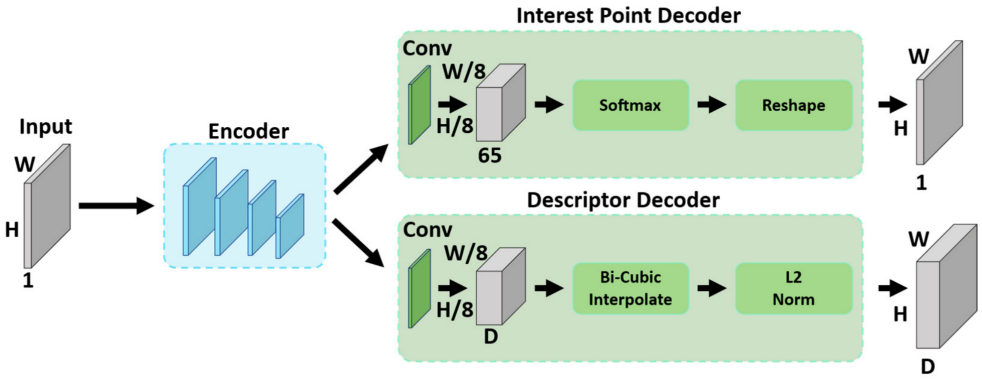



**Figure 5.** Architecture of the learning-based feature extractor, which consists of a shared encoder and two decoders.

For an input image of size $W \times H$, the feature detection branch first outputs a tensor of size $W/8 \times H/8 \times 65$, where the 65 channels correspond to non-overlapping $8 \times 8$ local grids plus an additional channel for non-feature points. Subsequently, a Softmax operation is applied along the channel dimension to suppress the non-feature responses, followed by a reshaping operation to recover a keypoint probability map at the original resolution of $W \times H$.

In the descriptor generation stage, SuperPoint adopts a semi-dense decoding approach rather than directly generating fully dense descriptors. This design balances parameter scale with operational efficiency. Subsequently, the system performs bicubic interpolation on the descriptors, followed by $L_2$ normalization to obtain the final local feature representation. As the SuperPoint module within the system framework, this method exhibits good adaptability under complex illumination changes, weak texture, and degraded scenes. Compared with traditional manually designed features, learning-based local features provide more stable keypoint responses and more discriminative description capability, thereby improving feature recognition and matching quality, and demonstrating robustness superior to that of some traditional feature-based visual SLAM methods in complex environments. More importantly, SuperPoint possesses strong cross-scene generalization capability and can adapt well to a variety of challenging real-world environments.

Figures 6 and 7 show a comparison of feature extraction results between the proposed method and ORB-SLAM3 in the EuRoC indoor low-light scenes and the 4Seasons outdoor low-light scenes, respectively. The left column shows the results of ORB-SLAM3, and the right column shows the results of the proposed method. It can be observed that in low-light regions, ORB-SLAM3 extracts only a limited number of feature points, and their distribution is noticeably biased toward areas with stronger local illumination. In dark regions, due to insufficient image brightness and loss of detail, the available features are sparse, and in some cases no usable feature points can be extracted at all. In contrast, the proposed method extracts a larger number of effective feature points with a more uniform distribution under the same low-light conditions. It maintains good feature coverage especially in dark and weak-texture regions where traditional methods struggle to respond. This demonstrates that the low-light enhancement and learning-based feature extraction framework constructed in this paper significantly improves front-end feature perception quality in both indoor and outdoor low-light environments, thereby providing a more stable foundation for subsequent feature matching, continuous tracking, and pose optimization.

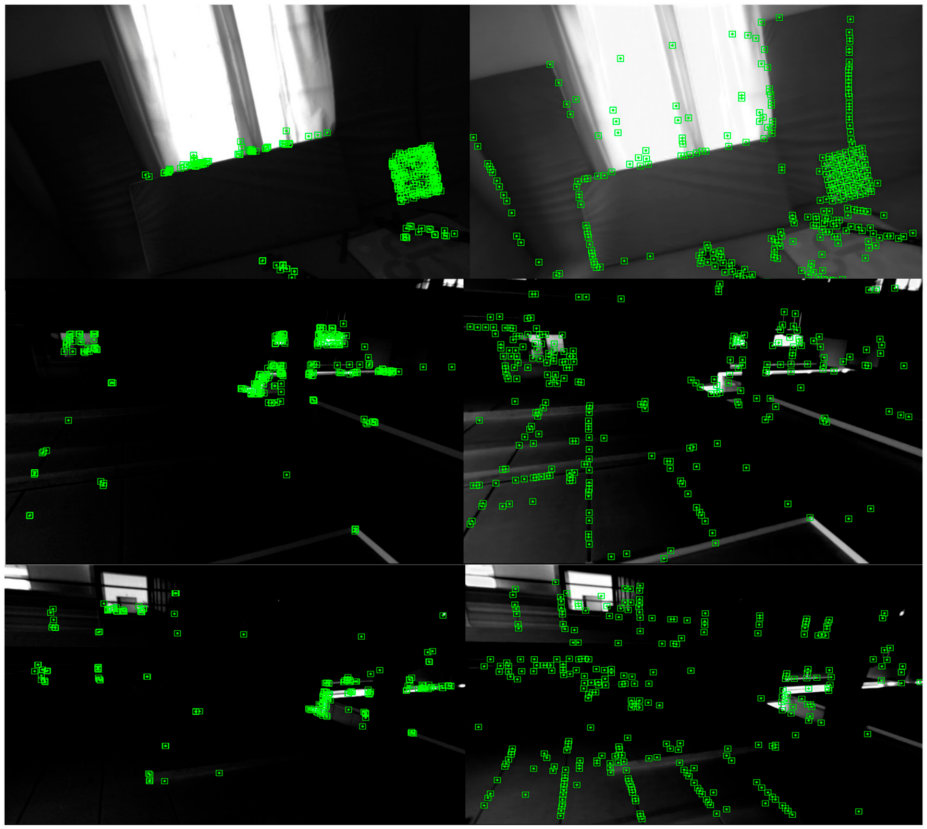



**Figure 6.** Comparison of feature extraction performance on the EuRoC dataset.

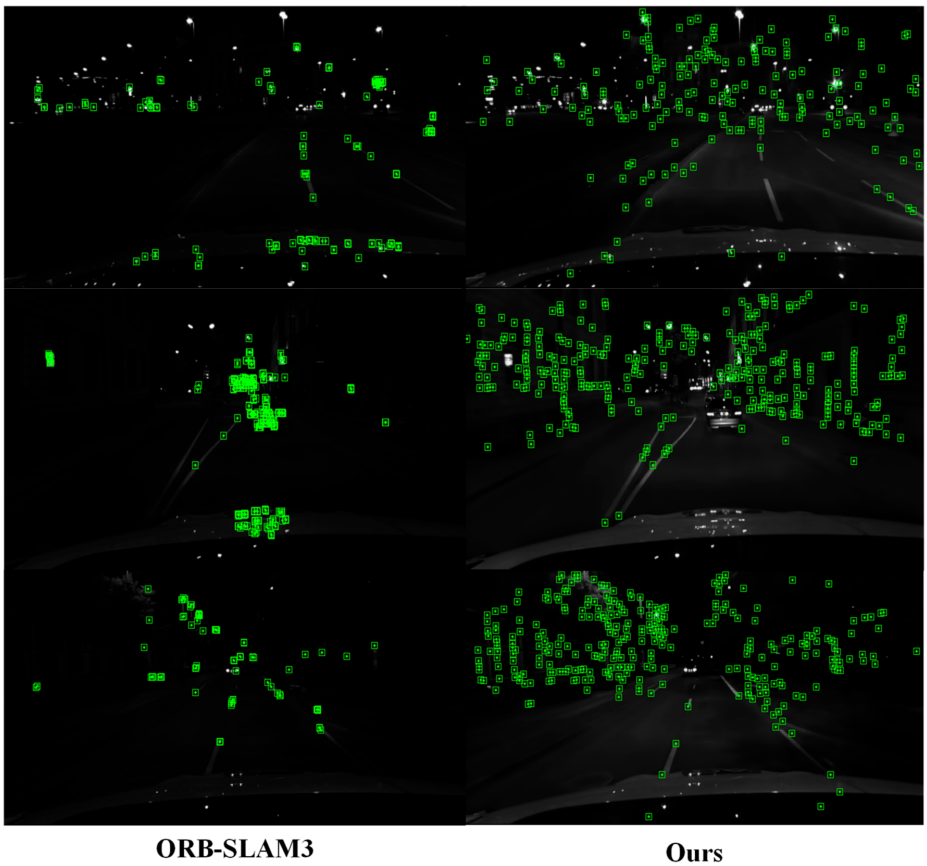



**Figure 7.** Comparison of feature extraction performance on the 4Seasons dataset.

### 3.4. Deep Local Feature-Matching Network

After obtaining local features from the enhanced images, this paper employs LightGlue to perform feature association between consecutive frames. LightGlue is an established matching network; therefore, this section summarizes only the elements needed for integration into RetiGlue-SLAM and refers readers to [41] for the full derivation. Given two sets of local features, LightGlue updates their hidden-state representations through self-attention and cross-attention layers and predicts cross-frame correspondences from descriptor similarity and contextual structure, as shown in Figure 8.

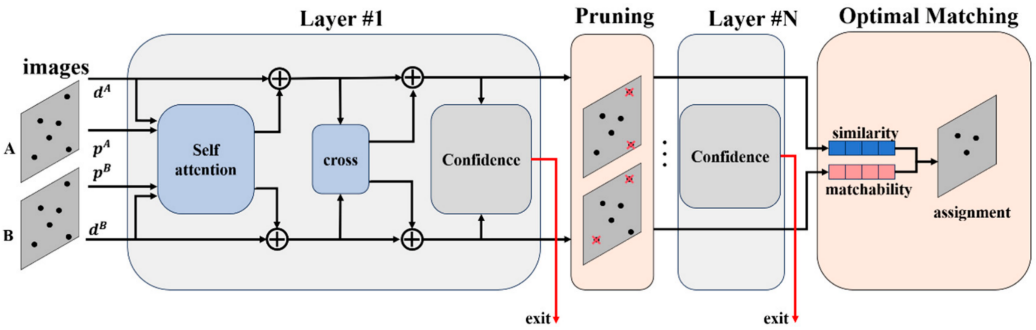



**Figure 8.** Workflow of the LightGlue local feature-matching network.

Let the two image frames be A and B, containing M and N local features, respectively. The feature sets are A = {1, . . ., M} and B = {1, . . ., N}, and the corresponding descriptor for feature i in image I is $d_i^l$.

$$A = \{1,\ldots,M\},\ B = \{1,\ldots,N\} \tag{16}$$

For image I, feature i is represented by a hidden state $x_i^l$. Its initial value is the input descriptor, $x_i^l = d_i^l$.

$$x_i^I \leftarrow d_i^I \tag{17}$$

The network iteratively updates the hidden states through self-attention and cross-attention layers. Each layer uses a multi-layer perceptron (MLP) to refine the current state using information aggregated from the other image, so the final state contains both local descriptor information and contextual constraints.

To characterize correspondences, LightGlue computes a similarity score for every feature pair. For feature i in image A and feature j in image B, the score is defined in Equation (18).

$$S_{ij} = Linear\left(x_i^A\right)^T Linear\left(x_j^B\right), (i,j) \in \mathcal{A} \times \mathcal{B} \tag{18}$$

Here, Linear(.) denotes a learnable linear transformation with bias. Since the score is computed after multi-layer interaction, it reflects the updated feature context rather than only the original descriptors.

The network also predicts a matchability score for each feature. For feature i, the score in Equation (19) indicates whether the feature is suitable for the final match.

$$\sigma_i = \text{Sigmoid}(\text{Linear}(x_i)) \in [0,1] \tag{19}$$

A score near zero indicates that no reliable match was found, whereas a score near one indicates high matching potential. This score suppresses unstable features before correspondences are established.

LightGlue then combines similarity and matchability into a joint assignment probability. For a feature pair (i, j), the probability is defined in Equation (20).

$$P_{ij} = \sigma_i^A \sigma_j^B \cdot Softmax_{k \in \mathcal{A}}(S_{kj})_i \cdot Softmax_{k \in \mathcal{B}}(S_{ik})_j \quad (20)$$

In Equation (20), sigma_iˆA and sigma_jˆB are the matchability scores in the two images. The final assignment is therefore constrained by both descriptor similarity and the reliability of the two features.

Accepted matches are forwarded as candidate correspondences to the tracking thread. The LightGlue stage itself does not perform pose estimation or map optimization; those operations remain part of the ORB-SLAM3 back end.

Figure 9 shows a comparison of feature-matching results between the proposed method and ORB-SLAM3 on the EuRoC dataset and the 4Seasons dataset. To ensure consistency in the comparison, the two images shown in each group are from the same scene, with a frame interval of 20 frames, reflecting matching stability under certain viewpoint changes and appearance variations. In each group of results, the left column shows the matching results of ORB-SLAM3, and the right column shows those of the proposed method. It can be observed that in low-light scenes, ORB-SLAM3 can establish a relatively limited number of valid matches, and the matching distribution is mainly concentrated in regions with higher local brightness. Meanwhile, a certain number of mismatches still exist in dark regions and areas with degraded details; for example, in the lower-left corner of the first result, noticeable incorrect correspondences can be seen. This indicates that traditional manually designed features are susceptible to insufficient brightness, increased noise, and texture degradation under low-illumination conditions, leading to insufficient matching constraints or even introducing incorrect associations.

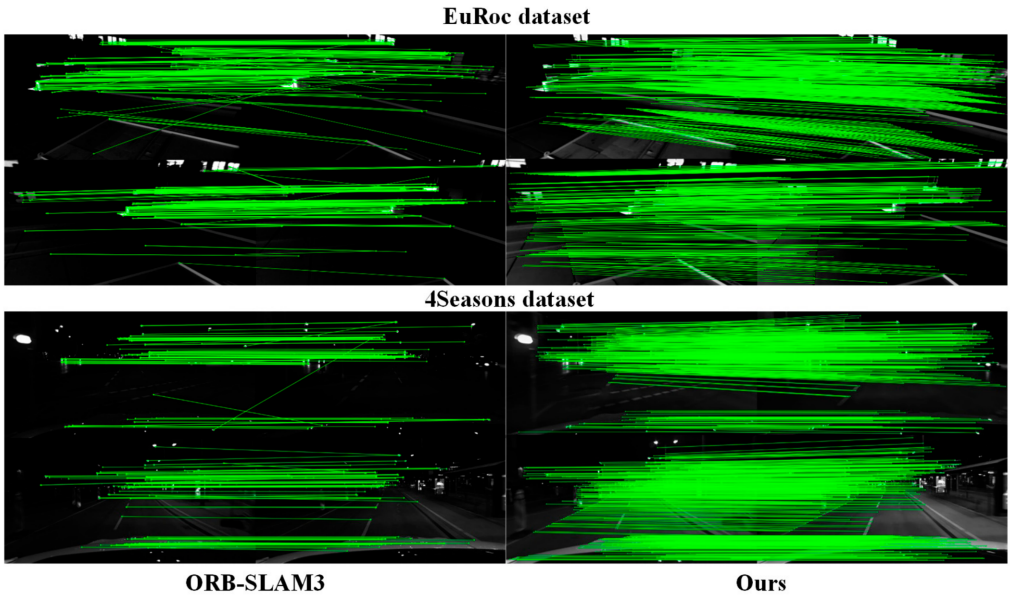



**Figure 9.** Comparison of matching performance on different datasets.

In contrast, the proposed method obtains a larger number of correctly matched points with a more uniform spatial distribution under the same scenes, while suppressing mismatches in low-light conditions. Retinexformer improves the observability of low-light images, SuperPoint provides more stable local representations, and LightGlue strengthens cross-frame association. These factors provide more reliable candidate correspondences to the tracking thread.

The coupling is intentionally limited to the front end: Retinexformer improves image quality, SuperPoint supplies local features, and LightGlue supplies candidate matches, whereas the geometric back end remains unchanged. This design makes the source of any localization improvement traceable to the modified observation pipeline.

### *3.5. Module Coupling and Implementation Details*

The three learned components are coupled through a fixed, sequential data flow rather than by retraining the full SLAM system. For each synchronized stereo pair, Retinexformer enhances the left and right images before feature extraction. SuperPoint then produces keypoints and descriptors for the enhanced images, and LightGlue matches descriptors between the current and previous frames. Only accepted correspondences are forwarded to the tracking thread; map updates, keyframe selection, IMU pre-integration, local bundle adjustment, and loop closure remain those of the unmodified ORB-SLAM3 back end. This separation keeps the geometric optimization transparent and allows the front-end modules to be replaced independently.

The prototype was implemented on Ubuntu 20.04 with an Intel Core i9-10900 CPU and an NVIDIA GeForce RTX 3090 GPU. Pretrained models supplied by the original authors were used for Retinexformer, SuperPoint, and LightGlue. ORB-SLAM3 parameters were kept at their default settings for each sequence, as described in Section 4.1. The learned modules run sequentially in the current prototype; consequently, the latency values in Section 4.5 measure front-end processing only and should not be interpreted as end-to-end SLAM throughput.

## 4. Results and Discussion

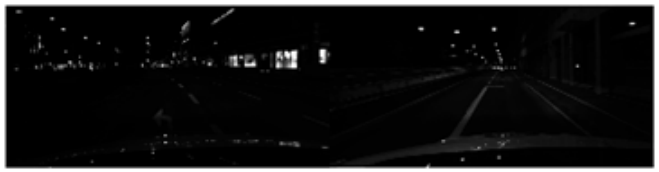
### *4.1. Datasets and Experimental Setup*

This section evaluates RetiGlue-SLAM on publicly available indoor and outdoor low-light sequences. The goal is to test front-end feature perception, tracking completion, and localization accuracy under underexposure and illumination variation; it is not a field-deployment study.

All experiments were conducted on a computer equipped with the Ubuntu 20.04 operating system, an Intel Core i9-10900 CPU, and a GeForce RTX 3090 GPU. The deep learning networks used pretrained models provided by the official sources. It is worth noting that the original open-source framework of ORB SLAM3 contains a set of core parameters that affect system performance, such as the number of image pyramid levels and scale factor, variables related to feature extraction and matching, camera frame rate, and depth threshold. To ensure comparability of the experimental results, this paper maintains the same parameter settings as the original ORB SLAM3 during testing on each dataset sequence. The experimental results show that the proposed method runs stably without requiring significant additional tuning of the underlying configuration, which indirectly demonstrates the system's good transferability and practical applicability.

To validate the feasibility of the experiments, we selected publicly available datasets that are widely used for evaluating SLAM localization accuracy. Table 1 presents the main characteristics of the datasets used in our experiments. According to the lighting conditions of different sequence scenes, we divided the datasets into two categories and conducted corresponding experimental tests to evaluate the robustness and accuracy of the system under indoor and outdoor low-light conditions.

**Table 1.** Main aspects related to the presented benchmark datasets.

| Dataset | Scene Type | Scene Characteristics | Time Period | Lighting Condition | Sequence Amount | Measurement Unit | Sample |
|---|---|---|---|---|---|---|---|
| EuRoC | Indoor | Room | Morning | Normal | 6 | cm |  |
| | | Large factory | Evening | Dusky | 5 | cm | |
| 4Seasons | Outdoor | City streets | Night | Dark | 5 | m |  |

https://doi.org/10.3390/ijgi15100468

We use low-light as an operational image-observation condition rather than as a universal photometric threshold: a sequence is treated as low-light when large image regions remain underexposed, contrast and texture are reduced, and ORB-SLAM3 feature tracking becomes sparse or unstable. By complex, we mean the combination of such illumination degradation with motion blur, rapid rotation, weak texture, or dynamic interference. The EuRoC MH_04 and MH_05 sequences and the night-time 4Seasons Old Town sequence represent these conditions in publicly available benchmarks. Normal-light EuRoC sequences are retained as controls. We do not claim that these benchmarks reproduce battlefield or field-deployment conditions; battlefield reconnaissance is only a long-term application motivation.

### *4.2. Results on the EuRoC Dataset*

To systematically evaluate the overall performance of the proposed method on standard visual SLAM benchmarks, this paper first conducts comprehensive experiments on the EuRoC dataset [42]. This dataset is one of the most commonly used public evaluation benchmarks in visual odometry and VSLAM research, covering a variety of indoor flight scenarios, and can comprehensively reflect the performance differences in algorithms under different environmental conditions. The EuRoC dataset was collected by ETH Zurich using a micro-aerial vehicle platform equipped with an MT9V034 stereo camera (20 Hz) and an ADIS16448 IMU (200 Hz). The scenes include two indoor rooms and a large industrial environment, providing a total of 11 stereo sequences, which encompass both sequences with favorable lighting and smooth motion, as well as high-speed motion sequences accompanied by motion blur and low illumination.

Given the significant differences in illumination conditions among the sub-scenes in EuRoC, this paper divides them into two categories and tests them separately. Among them, the industrial hall sequences MH_01 to MH_05 are classified as low-light scenes due to the relatively dark environment, as shown in Figure 10a, while the room sequences V1_01 to V2_03 have relatively sufficient lighting and are classified as well-lit scenes, as shown in Figure 10b Since this dataset provides both stereo images and IMU measurements, this paper further conducts experiments from three sensor configuration perspectives: stereo, monocular-inertial, and stereo-inertial modes. To quantify algorithm accuracy, the absolute trajectory error (ATE) is adopted as the evaluation metric, and the best results for each sequence are highlighted in bold. Considering that some internal processes of the system involve a degree of randomness, each set of experiments is repeated five times, and the median value is taken as the final result, which is then compared with other state-of-the-art SLAM methods, as shown in Table 2.

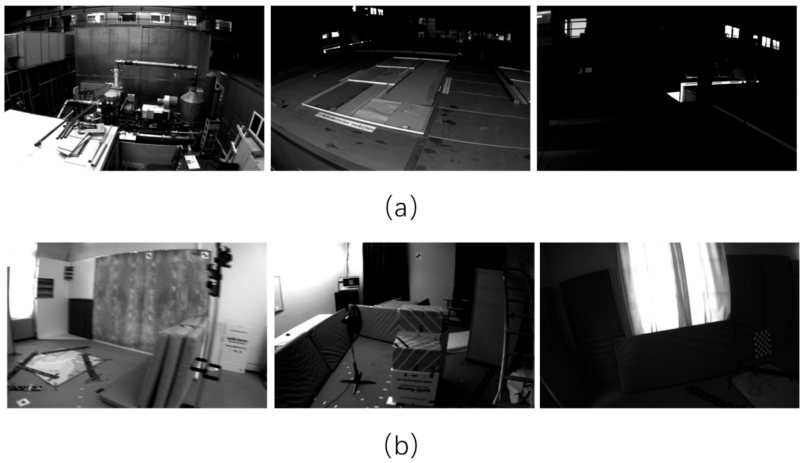



**Figure 10.** Scenes of the EuRoC dataset. (**a**) Machine Hall scene; (**b**) Vicon Room scene.

**Table 2.** Comparison of ATE RMSE (m) on the EuRoC dataset. * indicates that the corresponding method failed for that sequence; the best result is highlighted in bold.

| Dataset | Method | MH01 | MH02 | MH03 | MH04 | MH05 | V101 | V102 | V103 | V201 | V202 | V203 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mono + IMU | MSCKF [43] | 0.420 | 0.450 | 0.230 | 0.370 | 0.480 | 0.340 | 0.200 | 0.670 | 0.100 | 0.160 | 1.130 | 0.414 |
| | OKVIS [44] | 0.160 | 0.220 | 0.240 | 0.340 | 0.470 | 0.090 | 0.200 | 0.240 | 0.130 | 0.160 | 0.290 | 0.231 |
| | ROVIO [45] | 0.210 | 0.250 | 0.250 | 0.490 | 0.520 | 0.100 | 0.100 | 0.140 | 0.120 | 0.140 | 0.140 | 0.224 |
| | VINS-Mono [46] | 0.093 | 0.112 | 0.102 | 0.176 | 0.102 | 0.059 | 0.083 | 0.156 | 0.068 | 0.085 | 0.227 | 0.115 |
| | Kimera-VIO [47] | 0.110 | 0.100 | 0.160 | 0.240 | 0.350 | 0.050 | 0.080 | 0.070 | 0.080 | 0.100 | 0.210 | 0.141 |
| | ORB-SLAM3 | 0.044 | 0.061 | 0.052 | 0.073 | 0.063 | 0.046 | 0.017 | 0.031 | 0.049 | 0.022 | **0.040** | 0.045 |
| | **Ours** | **0.024** | **0.036** | **0.039** | **0.045** | **0.057** | **0.032** | **0.011** | **0.021** | **0.040** | **0.015** | 0.067 | **0.044** |
| Stereo + IMU | OKVIS | 0.134 | 0.083 | 0.191 | 0.191 | 0.249 | 0.046 | 0.077 | 0.097 | 0.059 | 0.096 | 0.114 | 0.117 |
| | VINS-Fusion | 0.163 | 0.154 | 0.280 | 0.280 | 0.287 | 0.077 | 0.069 | 0.113 | 0.063 | 0.091 | 0.096 | 0.138 |
| | Kimera | 0.080 | 0.090 | 0.150 | 0.150 | 0.240 | 0.050 | 0.110 | 0.120 | 0.070 | 0.100 | 0.190 | 0.119 |
| | BASALT [48] | 0.080 | 0.060 | 0.100 | 0.100 | 0.080 | 0.040 | 0.020 | 0.030 | 0.030 | 0.020 | * | 0.051 * |
| | ORB-SLAM3 | 0.033 | 0.037 | 0.051 | 0.051 | 0.076 | 0.041 | 0.013 | 0.024 | 0.036 | **0.014** | **0.022** | 0.033 |
| | OKVIS2[49] | 0.047 | 0.038 | 0.089 | 0.089 | 0.086 | 0.037 | 0.021 | 0.036 | 0.039 | 0.027 | 0.046 | 0.048 |
| | **Ours** | **0.014** | **0.025** | **0.024** | **0.043** | **0.035** | **0.032** | **0.009** | **0.018** | **0.019** | 0.015 | 0.045 | **0.025** |

Table 2 presents the comparison of ATE RMSE between the proposed method, ORB-SLAM3, and other state-of-the-art visual-inertial SLAM methods on the EuRoC dataset. Overall, the proposed method demonstrates competitive performance in both monocular-inertial and stereo-inertial configurations. In the monocular-inertial mode, the proposed method achieves the best or second-best results on most sequences, reducing the mean ATE RMSE from 0.045 m (ORB-SLAM3) to 0.044 m. Notably, on the MH04 and MH05 sequences, which are more challenging because of low illumination and complex motion, the errors are reduced from 0.073 m and 0.063 m to 0.045 m and 0.057 m, respectively. In the stereo-inertial mode, the mean ATE RMSE decreases from 0.033 m (ORB-SLAM3) to 0.025 m, with the proposed method achieving the best results on MH01, MH02, MH03, MH04, MH05, V101, V102, V103, and V201. On MH05, the error is reduced from 0.076 m to 0.035 m.

From the perspective of different scene types, the proposed method maintains good localization accuracy in both low-light industrial scenes and conventional indoor scenes. In low-light sequences such as MH04 and MH05, the advantage over the baseline is clearer. In normally illuminated sequences such as V101 and V201, the method also maintains low error, which indicates that the enhancement module does not necessarily impair performance under normal lighting. The proposed method does not achieve the best result on every sequence, so the claim is competitive accuracy rather than uniform superiority.

The two mean values in Table 2 correspond to two sensor configurations. The first is the mean over the Mono + IMU rows, and the second is the mean over the Stereo + IMU rows; they are not repeated estimates of the same experiment. In the Mono + IMU group, the mean improvement is small because the method is better on several sequences but worse on V2_03 (0.067 m versus 0.040 m for ORB-SLAM3). This underperformance is now discussed rather than hidden by the mean. It likely occurs because normal-light V2_03 does not benefit from low-light enhancement, while the learned front-end can produce a less favorable spatial distribution of keypoints than ORB features in that particular geometry. In the Stereo + IMU group, the mean decreases from 0.033 m to 0.025 m, and the improvement is more consistent. We therefore claim low-light robustness and competitive normal-light performance, not uniform superiority on every sequence.

### *4.3. Results on the 4Seasons Dataset*

To further validate the localization performance of the proposed method in outdoor low-light environments, this paper selects the publicly available 4Seasons dataset [50] for experimental evaluation, as shown in Figure 11. As a challenging benchmark for long-term visual localization, this dataset covers scenes spanning different seasons, weather conditions, and lighting variations, with significant changes in scene appearance, making it suitable for testing the robustness of SLAM systems under complex illumination conditions.



**Figure 11.** Scenes of the 4Seasons dataset.

Among them, the oldtown_2021-05-10_21-32-00 sequence corresponds to a night-time old town scene, captured at 21:32. The overall lighting along the entire trajectory is weak, with only a few local light sources, posing significant challenges for vision-based localization and mapping methods. This sequence contains 24,658 image frames and covers a total distance of approximately 5713 m. To more intuitively analyze the system's performance under night-time low-light conditions, this paper divides this long sequence into five representative sub-segments based on trajectory distribution and scene characteristics, as shown in Figure 12 In the figure, the light blue curve represents the full trajectory, while the segments marked in different colors correspond to the five selected test sub-segments, denoted as OT_01_easy, OT_02_medium, OT_03_medium, OT_04_difficult, and OT_05_difficult. These sub-segments cover varying lighting conditions, trajectory shapes, and scene complexity, and can be used to analyze the adaptability and robustness of the proposed method in night-time urban environments when facing factors such as low illumination, continuous turns, and local lighting variations. By evaluating the long sequence in segments, the performance differences in the algorithm under scenes of varying difficulty can be more clearly reflected.

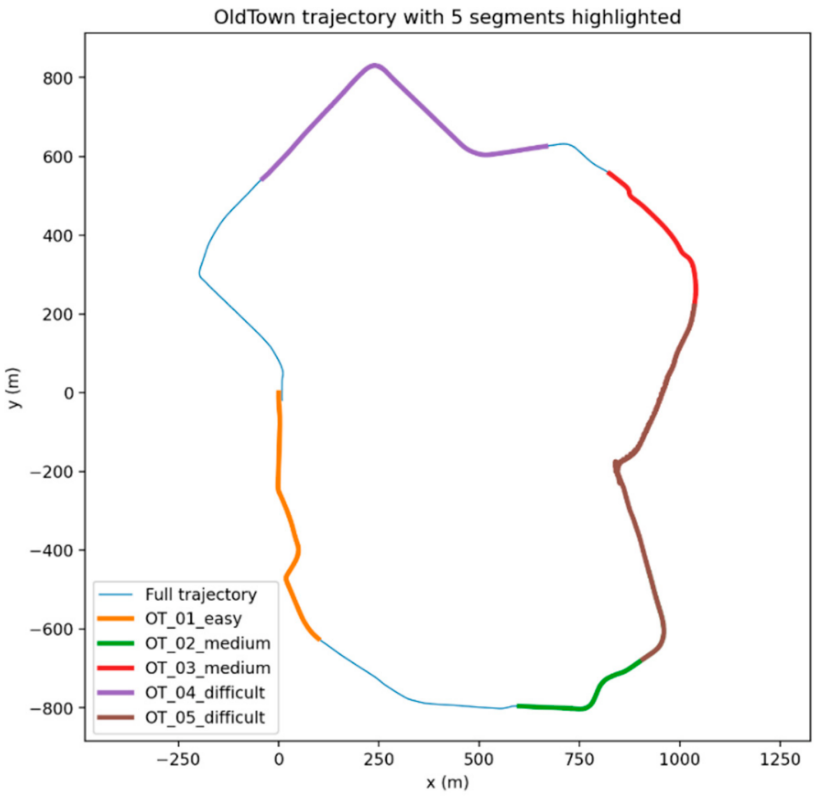



**Figure 12.** Schematic diagram of the segmentation of the night-time sequence in the Old Town scene of the 4Seasons dataset.

Table 3 presents a comparison of the translational RMSE between the proposed method, ORB-SLAM3, and ORB-SLAM2 on the night-time sequence of the Old Town scene in the 4Seasons dataset. The proposed method achieves the lowest error on all five sub-segments and is the only algorithm that successfully completes localization on every test sequence, with a success rate of 100%, compared with 80% for ORB-SLAM3 and 20% for ORB-SLAM2. On OT_01_easy and OT_02_medium, the proposed method reduces the RMSE to 29.063 m and 11.463 m, corresponding to decreases of approximately 30.46% and 25.82% relative to ORB-SLAM3. On OT_03_medium, the error is reduced from 22.475 m to 15.899 m, a reduction of about 29.26%. On OT_04_difficult, ORB-SLAM3 fails, whereas the proposed method estimates the full trajectory. On OT_05_difficult, the proposed method reduces the error from 34.975 m to 6.857 m.

**Table 3.** Comparison of translation RMSE (m) on the 4Seasons dataset. X indicates tracking failure.

| Sequence | Scene Characteristics | Ours | ORB-SLAM3 (Stereo) | ORB-SLAM2 (Stereo) |
|---|---|---|---|---|
| OT_01_easy | Reduced lighting at the corner | 29.063 | 41.793 | X |
| OT_02_medium | Low lighting and continuous turns | 11.463 | 15.452 | 17.351 |
| OT_03_medium | Sudden darkening of lighting | 15.899 | 22.475 | X |
| OT_04_difficult | Limited streetlights and few structured objects | 45.896 | X | X |
| OT_05_difficult | Significantly degraded lighting environment | 6.857 | 34.975 | X |
| Success rate | - | 100% | 80% | 20% |

Figure 13 further presents a comparison of localization results between the proposed method and ORB-SLAM3 on the five night-time sub-segments of the 4Seasons dataset in the form of trajectories, where the red dashed line indicates the ground truth trajectory, and the blue solid line and yellow solid line represent the estimated trajectories of the proposed method and ORB-SLAM3, respectively. As shown in Figure 13a–c, on the OT_01_easy, OT_02_medium, and OT_03_medium sequences, the estimated trajectory of the proposed method is overall closer to the ground truth compared to that of ORB-SLAM3. Notably, in regions involving continuous turns, locally insufficient lighting, and sudden brightness changes, the proposed method maintains good trajectory consistency. In contrast, ORB-SLAM3 is more prone to accumulated drift in dark areas and turning sections, resulting in noticeable deviations between the estimated trajectory and the true path.

For the more challenging OT_04_difficult and OT_05_difficult sequences, as shown in Figure 13d,e, the traditional method exhibits more severe drift in scenes with even weaker illumination and less reliable structural information, and in some cases fails to complete stable tracking throughout the entire trajectory. In contrast, the proposed method still produces a complete trajectory and maintains good consistency with the ground truth in terms of overall path shape. These results indicate that the proposed RetiGlue-SLAM provides more stable front-end observations and data association constraints in night-time low-light environments, effectively suppressing trajectory drift and enhancing the reliability of the system for localization in complex outdoor low-light scenes.

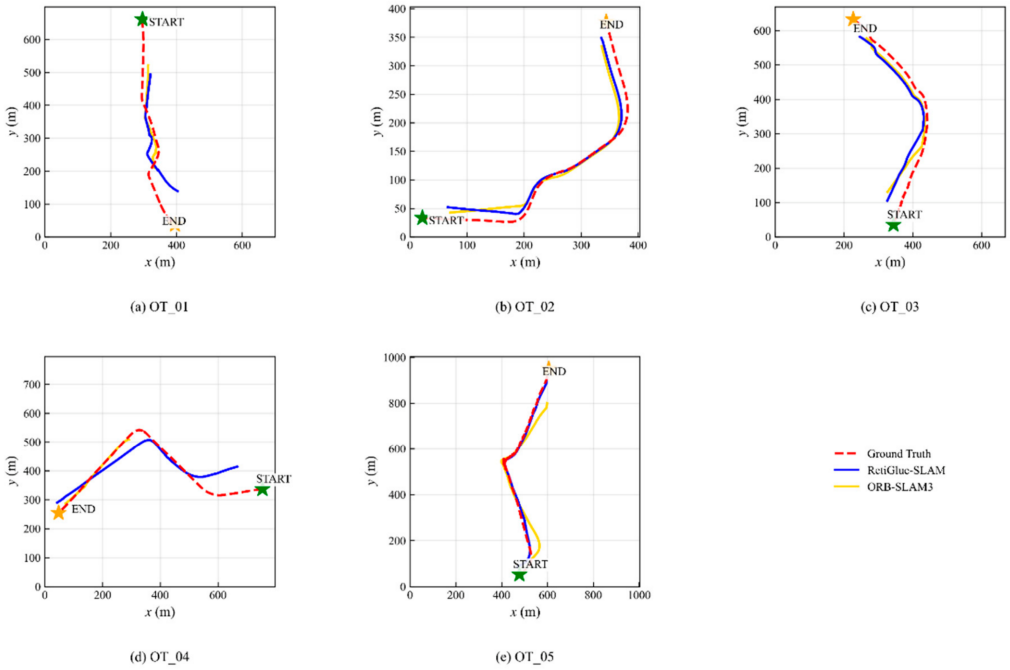



**Figure 13.** Comparison of trajectory estimation results between the proposed method and ORB-SLAM3 with the ground truth.

### *4.4. Ablation Study*

To examine whether the improvement comes only from the complete pipeline, we conducted an ablation study on three representative EuRoC sequences: MH_04, MH_05, and V1_03. The compared configurations are (i) the original ORB-SLAM3 front end, (ii) SuperPoint with LightGlue without Retinexformer, and (iii) the full RetiGlue-SLAM system. All other settings were kept identical. Table 4 reports the ATE RMSE results.

**Table 4.** Ablation study of the learned front-end and the full RetiGlue-SLAM pipeline on representative EuRoC sequences (ATE RMSE, m).

| Sequence | ORB-SLAM3 | SuperPoint + LightGlue | RetiGlue-SLAM (Ours) | Improvement vs. ORB-SLAM3 (%) |
|---|---|---|---|---|
| MH04 | 0.073 | 0.059 | 0.045 | 38.4 |
| MH05 | 0.063 | 0.060 | 0.057 | 9.5 |
| V103 | 0.031 | 0.035 | 0.021 | 32.3 |

SuperPoint + LightGlue reduces the ATE on MH_04 and MH_05 from 0.073 m and 0.063 m to 0.059 m and 0.060 m, but it increases the error on V1_03 from 0.031 m to 0.035 m. Adding Retinexformer reduces the errors to 0.045 m, 0.057 m, and 0.021 m, respectively, corresponding to reductions of 38.4%, 9.5%, and 32.3% relative to ORB-SLAM3. The V1_03 result is important: learned matching alone is not sufficient and can be sensitive to a specific scene, whereas the enhancement-coupled pipeline provides the most consistent improvement. Because this ablation does not isolate Retinexformer alone or every possible module combination, we interpret the result as evidence of the benefit of the full coupling rather than as a universal claim that one module dominates in all environments.

### *4.5. Runtime Analysis*

The front-end latency of the current single-threaded prototype was measured on the same platform described in Section 4.1 using 752 × 480 stereo imagery and 500 keypoints per image. Table 5 reports the mean processing time for each learned component. Retinex-

former processes the stereo pair in approximately 27 ms, SuperPoint requires 8 ms for feature extraction, and LightGlue requires 6 ms for matching. The total front-end latency is therefore approximately 41 ms per stereo frame, corresponding to 24.39 frames per second (FPS).

**Table 5.** Module-level front-end latency of the RetiGlue-SLAM prototype.

| Module | Mean Latency (ms) | Share of Front-End Latency (%) |
|---|---|---|
| Retinexformer | 27 | 65.9 |
| SuperPoint | 8 | 19.5 |
| LightGlue | 6 | 14.6 |
| Total front end | 41 | 100.0 |

These values describe only the learned front-end pipeline. They exclude feature management, local map tracking, keyframe insertion, local bundle adjustment, loop closure, and global optimization, which remain part of ORB-SLAM3. Consequently, 41 ms is a lower bound on the complete system latency, and the measured 24.39 FPS is not an end-to-end real-time guarantee. We report it to make the computational cost transparent and to identify Retinexformer as the dominant front-end cost. Future work will parallelize enhancement and feature processing and evaluate end-to-end runtime.

### *4.6. Comparison with Recent Low-Light-Oriented SLAM Methods*

To relate our results to very recent low-light-oriented systems, we compare design choices rather than raw numbers. Light-SLAM [16] also uses LightGlue and reports robust performance under challenging lighting, but it targets a different front-end and back-end configuration and a different benchmark suite. The unsupervised enhancement method in [17] focuses on construction-site localization, BVT-SLAM [18] requires a visible–thermal sensor pair, and LDFE-SLAM [19] uses a light-aware deep front-end. These systems are not directly comparable with the EuRoC and 4Seasons evaluation in this paper because of differences in sensor modality, training data, benchmark sequences, and code availability. A controlled numerical comparison would require reimplementing each method on identical inputs and hardware. We therefore restrict quantitative claims to ORB-SLAM3 and the baselines shown in Tables 2 and 3, while using [16–19] to position the proposed design. RetiGlue-SLAM is closest in spirit to [16], but differs by coupling enhancement and learned matching with an unchanged geometric back-end and by reporting a component-level ablation and transparent front-end latency. Within the enhancement stage, conventional histogram-equalization methods [26] are computationally simple but often amplify noise, whereas domain-specific unsupervised enhancement [17] requires suitable unlabeled data. Retinexformer [37] was selected because it jointly models illumination and degradation in one stage and provides publicly available pretrained weights. This is a design-level justification; an empirical comparison of alternative enhancers under matched training and tuning budgets was not available in this study.

### *4.7. Strengths, Weaknesses, and Limitations*

The main strength of RetiGlue-SLAM is that it improves low-light observations without discarding the well-tested geometric back-end of ORB-SLAM3. The ablation and runtime analysis clarify both the source of the improvement and its computational cost. The main limitations are equally explicit. First, the evaluation uses public benchmark sequences rather than a newly collected field dataset. Second, this ablation does not isolate every component or compare alternative enhancement networks under identical training

and tuning budgets. Third, the front-end latency is not equivalent to end-to-end real-time performance. Fourth, the method does not outperform the baseline on every normal-light sequence, as shown by V2_03 in the Mono + IMU configuration. These limitations define the scope of the claims: the method is supported as a robust low-light benchmark solution, not as a universally superior or field-proven deployment system.

## 5. Conclusions and Future Work

This paper presented RetiGlue-SLAM, a low-light visual-inertial SLAM framework that retains ORB-SLAM3 as the geometric back end and modifies the front-end observation pipeline with Retinexformer, SuperPoint, and LightGlue.

Experiments on low-light EuRoC sequences and the night-time 4Seasons Old Town sequence showed improved feature-matching, tracking completion, and localization accuracy relative to the baseline. The two-stage ablation indicated that the full coupling provides more consistent improvement than learned matching alone, while the runtime analysis identified Retinexformer as the dominant front-end cost.

The present study is limited to publicly available benchmark sequences, does not include a newly collected field dataset, and does not establish end-to-end real-time operation. Future work will therefore evaluate alternative enhancement networks under matched training and tuning budgets, parallelize the front-end modules and measure complete system latency, perform a full factorial ablation, and collect a representative field dataset before making deployment claims.

**Author Contributions:** Conceptualization, Yiming Li; methodology, Yiming Li and Xianpu Liang; software, Xianpu Liang; validation, Xianpu Liang; formal analysis, Xianpu Liang; investigation, Guangming Guo and Qi An; resources, Peng Su; data curation, Xianpu Liang; writing—original draft preparation, Xianpu Liang; writing—review and editing, Yiming Li and Peng Su; visualization, Xianpu Liang; supervision, Yiming Li; project administration, Yiming Li; funding acquisition, Guangming Guo and Qi An. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Young Backbone Teachers Support Plan of Beijing Information Science and Technology University, grant no. YBT 202405.

**Data Availability Statement:** The EuRoC and 4Seasons datasets used in this study are publicly available from their official sources. The EuRoC MAV dataset is available at https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets (accessed on 10 October 2024); the 4Seasons dataset is available at https://www.4seasons-dataset.com/ (accessed on 10 October 2024).

**Acknowledgments:** The authors would like to thank the providers of the EuRoC and 4Seasons datasets.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19929–19953. [CrossRef]
2. Lee, K.-H.; Hwang, J.-N.; Okopal, G.; Pitton, J. Ground-Moving-Platform-Based Human Tracking Using Visual SLAM and Constrained Multiple Kernels. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3602–3612. [CrossRef]
3. Shen, Z.; He, Y.; Du, X.; Yu, J.; Wang, H.; Wang, Y. YCANet: Target Detection for Complex Traffic Scenes Based on Camera-LiDAR Fusion. *IEEE Sens. J.* **2024**, *24*, 8379–8389. [CrossRef]
4. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef]
5. Klein, G.; Murray, D. Parallel Tracking and Mapping on a Camera Phone. In Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 19–22 October 2009; pp. 83–86. [CrossRef]
6. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]

7. Mur-Artal, R.; Tardos, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
8. Campos, C.; Elvira, R.; Rodriguez, J.J.G.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
9. Hu, K.; Wu, J.; Li, Y.; Lu, M.; Weng, L.; Xia, M. FedGCN: Federated Learning-Based Graph Convolutional Networks for Non-Euclidean Spatial Data. *Mathematics* **2022**, *10*, 1000. [CrossRef]
10. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 722–739. [CrossRef]
11. Feng, D.; Haase-Schutz, C.; Rosenbaum, L.; Hertlein, H.; Glaser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1341–1360. [CrossRef]
12. Li, R.; Wang, S.; Long, Z.; Gu, D. UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 7286–7291. [CrossRef]
13. Wang, S.; Clark, R.; Wen, H.; Trigoni, N. DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 2043–2050. [CrossRef]
14. Teed, Z.; Deng, J. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *arXiv* **2021**, arXiv:2108.10869. [CrossRef]
15. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236. [CrossRef]
16. Zhao, Z.; Wu, C.; Kong, X.; Li, Q.; Guo, Z.; Lv, Z.; Du, X. Light-SLAM: A Robust Deep-Learning Visual SLAM System Based on LightGlue Under Challenging Lighting Conditions. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 9918–9931. [CrossRef]
17. Chen, X.; Yu, Y. An Unsupervised Low-Light Image Enhancement Method for Improving V-SLAM Localization in Uneven Low-Light Construction Sites. *Autom. Constr.* **2024**, *162*, 105404. [CrossRef]
18. Qin, L.; Wu, C.; Kong, X.; You, Y.; Zhao, Z. BVT-SLAM: A Binocular Visible-Thermal Sensors SLAM System in Low-Light Environments. *IEEE Sens. J.* **2024**, *24*, 11599–11609. [CrossRef]
19. Liu, C.; Wang, Y.; Luo, W.; Peng, Y. LDFE-SLAM: Light-Aware Deep Front-End for Robust Visual SLAM Under Challenging Illumination. *Machines* **2026**, *14*, 44. [CrossRef]
20. Harris, C.; Stephens, M. A Combined Corner and Edge Detector. In Proceedings of the Alvey Vision Conference 1988, Manchester, UK, 31 August–2 September 1988; pp. 23.1–23.6. [CrossRef]
21. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
22. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In *Computer Vision—ECCV 2006*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3951, pp. 404–417. [CrossRef]
23. Rosten, E.; Porter, R.; Drummond, T. Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 105–119. [CrossRef]
24. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571. [CrossRef]
25. Labbe, M.; Michaud, F. RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation. *J. Field Robot.* **2019**, *36*, 416–446. [CrossRef]
26. Fang, Y.; Shan, G.; Wang, T.; Li, X.; Liu, W.; Snoussi, H. HE-SLAM: A Stereo SLAM System Based on Histogram Equalization and ORB Features. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 4272–4276. [CrossRef]
27. Shamwell, E.J.; Leung, S.; Nothwang, W.D. Vision-Aided Absolute Trajectory Estimation Using an Unsupervised Deep Network with Online Error Correction. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 2524–2531. [CrossRef]
28. Zhai, Y.; Liu, L.; Zhang, L.; Liu, Y. PoseConvGRU: A Monocular Approach for Visual Ego-Motion Estimation by Learning. *arXiv* **2019**, arXiv:1906.08095. [CrossRef]
29. Xiu, H.; Liang, Y.; Zeng, H.; Li, Q.; Liu, H.; Fan, B.; Li, C. Robust Self-Supervised Monocular Visual Odometry Based on Prediction-Update Pose Estimation Network. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105481. [CrossRef]
30. Xin, Z.; Wang, Z.; Yu, Z.; Zheng, B. ULL-SLAM: Underwater Low-Light Enhancement for the Front-End of Visual SLAM. *Front. Mar. Sci.* **2023**, *10*, 1133881. [CrossRef]

31. Parisotto, E.; Chaplot, D.S.; Zhang, J.; Salakhutdinov, R. Global Pose Estimation with an Attention-Based Recurrent Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 237–246. [CrossRef]
32. Bruno, M.S.; Colombini, E.L. LIFT-SLAM: A Deep-Learning Feature-Based Monocular Visual SLAM Method. *Neurocomputing* **2021**, *455*, 97–110. [CrossRef]
33. Tateno, K.; Tombari, F.; Laina, I.; Navab, N. CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6565–6574. [CrossRef]
34. Tang, J.; Folkesson, J.; Jensfelt, P. Geometric Correspondence Network for Camera Motion Estimation. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1010–1017. [CrossRef]
35. Tang, J.; Ericson, L.; Folkesson, J.; Jensfelt, P. GCNv2: Efficient Correspondence Prediction for Real-Time SLAM. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3505–3512. [CrossRef]
36. Qin, Z.; Yin, M.; Li, G.; Yang, F. SP-Flow: Self-Supervised Optical Flow Correspondence Point Prediction for Real-Time SLAM. *Comput. Aided Geom. Des.* **2020**, *82*, 101928. [CrossRef]
37. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12470–12479. [CrossRef]
38. Song, W.; Suganuma, M.; Liu, X.; Shimobayashi, N.; Maruta, D.; Okatani, T. Matching in the Dark: A Dataset for Matching Image Pairs of Low-Light Scenes. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6009–6018. [CrossRef]
39. Singh, S.P.; Mazotti, B.; Rajani, D.M.; Mayilvahanan, S.; Li, G.; Ghaffari, M. Twilight SLAM: Navigating Low-Light Environments. *arXiv* **2023**, arXiv:2304.11310. [CrossRef]
40. Jiang, Y.; Zhu, C.; Yang, J. Research on Low-Light Environment SLAM System via Retinex Enhancement and Deep Feature Optimization. *IEEE Sens. J.* **2025**, *25*, 29992–30004. [CrossRef]
41. Lindenberger, P.; Sarlin, P.-E.; Pollefeys, M. LightGlue: Local Feature Matching at Light Speed. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17581–17592. [CrossRef]
42. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC Micro Aerial Vehicle Datasets. *Int. J. Robot. Res.* **2016**, *35*, 1157–1163. [CrossRef]
43. Mourikis, A.I.; Roumeliotis, S.I. A Multi-State Constraint Kalman Filter for Vision-Aided Inertial Navigation. In Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, 10–14 April 2007; pp. 3565–3572. [CrossRef]
44. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-Based Visual-Inertial Odometry Using Nonlinear Optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [CrossRef]
45. Bloesch, M.; Burri, M.; Omari, S.; Hutter, M.; Siegwart, R. Iterated Extended Kalman Filter Based Visual-Inertial Odometry Using Direct Photometric Feedback. *Int. J. Robot. Res.* **2017**, *36*, 1053–1072. [CrossRef]
46. Qin, T.; Li, P.; Shen, S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
47. Rosinol, A.; Abate, M.; Chang, Y.; Carlone, L. Kimera: An Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 1689–1696. [CrossRef]
48. Usenko, V.; Demmel, N.; Schubert, D.; Stuckler, J.; Cremers, D. Visual-Inertial Mapping with Non-Linear Factor Recovery. *IEEE Robot. Autom. Lett.* **2020**, *5*, 422–429. [CrossRef]
49. Hu, K.; Zhang, X.; Li, K.; Wu, K.; Dong, R. EM-LSD-Based Visual-Inertial Odometry with Point-Line Feature. *IEEE Sens. J.* **2023**, *23*, 30794–30804. [CrossRef]
50. Wenzel, P.; Wang, R.; Yang, N.; Cheng, Q.; Khan, Q.; von Stumberg, L.; Zeller, N.; Cremers, D. 4Seasons: A Cross-Season Dataset for Multi-Weather SLAM in Autonomous Driving. *arXiv* **2020**, arXiv:2009.06364. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.